Cybersecurity Threat Detection using Machine Learning and Deep Learning Techniques

Sudhakar
sudhak82_scs@jnu.ac.in
School of Computer and Systems Sciences, Jawaharlal Nehru University
New Delhi, India

Abstract

The rapid increase in internet-connected devices due to the implementation of the Internet of Things (IoT) and Industry 4.0 poses a significant challenge for cybersecurity threat detection systems to effectively detect all the malicious programs and events in the network. The threat landscape is also evolving in all types of attacks: botnet, malware, fileless malware, or intrusion. A learning detection system is required to detect malicious events by analyzing the behavioural pattern of the program. In this context, we have proposed models to identify the malicious programs and events in the system using machine learning and deep learning techniques. A deep learning approach for malware classification with fine-tune convolution neural networks (MCFT-CNN) using traditional and transfer learning has been proposed to classify the malware into their malware family. The proposed model reported 99.18% accuracy and 5.14ms prediction time on the MalImg dataset. A machine learning-based (ML-IDS) model has been proposed to classify the intrusions on the network. The proposed model has reported 99.51% accuracy in multiclass classification and 99.86% in binary class classification. Efficient incident handling and response process model has been proposed in the case of fileless malware. A lightweight machine learning model was proposed to detect botnet infections in the IoT networks. The model was fine-tuned with hyper-parameters and trained using early stopping to avoid overfitting. The proposed model shows 100% accuracy with 0.01% true-negative and 99.99% true-positive.

Keywords: Cybersecurity, Malware classification, Intrusion detection, Botnet detection, Machine learning, Deep learning

AI-ML Systems, October 21–24, 2021, BANGALORE, INDIA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

ACM Reference Format:
Sudhakar. 2021. Cybersecurity Threat Detection using Machine Learning and Deep Learning Techniques. In *Proceedings of First International Conference on AI-ML Systems (AI-ML Systems).* ACM, Banglore, India, 3 pages.

1 Introduction

The prominent threats in cybersecurity are malware, fileless malware, botnet, and intrusions. The malware is a malicious program that do things which any programs are not supposed to do in the system. It comprises all sort of malware like Adware, Trojans, Bots, Worm, Backdoor, and Fileless malware [14]. The fileless malware can compromise a system without making any changes or minor changes in the file system living entirely in the main memory, rootkit, or registry. The fileless malware still has all the capabilities like traditional malware making it hard to detect and more harmful [12]. Bot is a small malicious program that can target the vulnerable system to compromise and make it a part of larger botnet (Ro**bot net**work) controlled by a bot-master. A bot-master can launch a cyberattack using the botnet like sending spam, data theft, compromising confidential information, and distributed denial of service (DDoS) [11]. Further, intrusion can be classified into any kind of unauthorised or malicious activities in the network like network flood attack, malware attached with messages, and man-in-middle attack [13].

It is very crucial for any network to neutralize the cyberthreats in order to safeguard their infrastructure. We have used machine learning and deep learning techniques to detect and classify the cyberthreats. We required benchmark datasets with properties like diversified attacks and imbalance distribution of classes to train our models. Each models required fine-tuning of hyper-parameters to train efficiently. Our models have successfully classify the above cyberthreats. The contribution of the thesis is shown in figure 1.

2 Motivation

Cybercriminals use malware and other attack vectors to compromise vulnerable machines. The conventional machine learning malware detection and classification algorithms uses static features of malware for the training purpose. The features extracted by the static analysis is text-based, i.e. signature [7], Opcode sequence [9], control flow graph [4],

bytecode [3], and n-gram [9]; therefore, only the subset of malware sample data is included in the training process. Hence, it will degrade the accuracy of the machine learning or deep learning algorithms and is time-consuming compared to the approach that uses complete information of malware samples in the form of visual features.

Various machine learning and deep learning methods such as M-CNN [5], NSGA-II [2], Deep CNN [10], CNN BiGRU [16], IMCFN [15] and CapsNet [1] have been used in the literature to detect malware using visual features. The machine learning algorithms are required to process malware datasets and the inevitable work of features engineering. At the same time, deep learning shows promising results to classify malware images [1, 2, 5, 10, 15, 16].

Thus, we propose a novel MCFT-CNN model to address the above issues and trained with only visual features. We have achieved 99.18% accuracy and 5.14ms prediction time on the MalImg dataset [6]. The model shows significant improvement over a larger dataset (Microsoft Malware Challenge [8]) with 98.63% accuracy and 5.15ms prediction time. Our model performs significantly better than the existed state-of-art [14]. Similarly, in intrusion detection and botnet detection, we have used machine learning algorithms to efficiently classify intrusions and botnet attacks. We have also proposed an incident handling and response process in case of a fileless malware attack to analyze the attack and behaviour of the fileless malware. The proposed models perform significantly better than other models available in the literature [1–5, 7, 9, 10, 15, 16].

Our main contributions to the cybersecurity research based on machine learning and deep learning are listed in the following points-

- A novel deep learning model has been proposed to classify the malware using visual features without feature engineering and prior knowledge of binary code analysis or reverse engineering.
- A novel investigative model of incident handling and response has been proposed, especially in fileless malware. The model includes all the phases with memory forensic, analysis and investigation of such incidents.
- A machine learning model has been proposed to classify web-intrusion attacks. The model uses a univariate feature selection technique on the intrusion dataset (CIC-IDS2017).
- A lightweight machine learning model has been proposed to classify botnet attacks in IoT networks.

3 Thesis timeline and author contribution

We have published/communicated our proposed models for cybersecurity threat detection like malware classification (MCFT-CNN) [14], a process model to handle fileless malware [12], a model for intrusion detection (ML-IDS) [13] [under review], a survey on botnet detection and research challenges [11], and an anomaly-based botnet detection using machine learning model in IoT networks (ABBDIoT) [Communicated]. In this thesis, the PhD candidate completes the conceptualization of the problem, implementation, formal analysis, methodology formulation, validation, and writing. Supervisor: supervises the work to improve the quality, investigate and validate the methods applied, and reviewed the drafts to improve it. The thesis writing is in progress, expected to complete the first and second chapters by October 2021 and submit the PhD thesis in March 2022.

4 Conclusions and future plans

The proposed models successfully identified the cybersecurity attack vectors like malware, botnet, intrusion and fileless malware. These attack vectors play a crucial role to compromise the vulnerable machine and launch a full-scale attack. Future plans are to identify the machine susceptible to malware infection by studying the different parameters of individual machines and applying appropriate machine learning algorithms with suitable feature engineering method.

References

[1] Aykut Çayır, Uğur Ünal, and Hasan Dağ. 2021. Random CapsNet forest model for imbalanced malware type classification task. *Computers & Security* 102 (2021), 102133.

[2] Zhihua Cui, Lei Du, Penghong Wang, Xingjuan Cai, and Wensheng Zhang. 2019. Malicious code detection based on CNNs and multi-objective algorithm. *J. Parallel and Distrib. Comput.* 129 (2019), 50–58.

[3] Jake Drew, Michael Hahsler, and Tyler Moore. 2017. Polymorphic malware detection using sequence classification methods and ensembles. *EURASIP Journal on Information Security* 2017, 1 (2017), 1–12.

[4] Mehadi Hassen and Philip K Chan. 2017. Scalable function call graph-based malware classification. In *Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy*. 239–248.

[5] Mahmoud Kalash, Mrigank Rochan, Noman Mohammed, Neil DB Bruce, Yang Wang, and Farkhund Iqbal. 2018. Malware classification with deep convolutional neural networks. In *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*. IEEE, 1–5.

[6] Lakshmanan Nataraj, Sreejith Karthikeyan, Gregoire Jacob, and Bangalore S Manjunath. 2011. Malware images: visualization and automatic classification. In *Proceedings of the 8th international symposium on visualization for cyber security*. 1–7.

[7] Edward Raff and Charles Nicholas. 2018. Lempel-Ziv Jaccard Distance, an effective alternative to ssdeep and sdhash. *Digital Investigation* 24 (2018), 34–49.

[8] Royi Ronen, Marian Radu, Corina Feuerstein, Elad Yom-Tov, and Mansour Ahmadi. 2018. Microsoft malware classification challenge. *arXiv preprint arXiv:1802.10135* (2018).

[9] Asaf Shabtai, Robert Moskovitch, Clint Feher, Shlomi Dolev, and Yuval Elovici. 2012. Detecting unknown malicious code by applying classification techniques on opcode patterns. *Security Informatics* 1, 1 (2012), 1–22.

[10] Ajay Singh, Anand Handa, Nitesh Kumar, and Sandeep Kumar Shukla. 2019. Malware classification using image representation. In *International Symposium on Cyber Security Cryptography and Machine Learning*. Springer, 75–92.

[11] Sudhakar and Sushil Kumar. 2019. Botnet Detection Techniques and Research Challenges. In *2019 International Conference on Advances in*

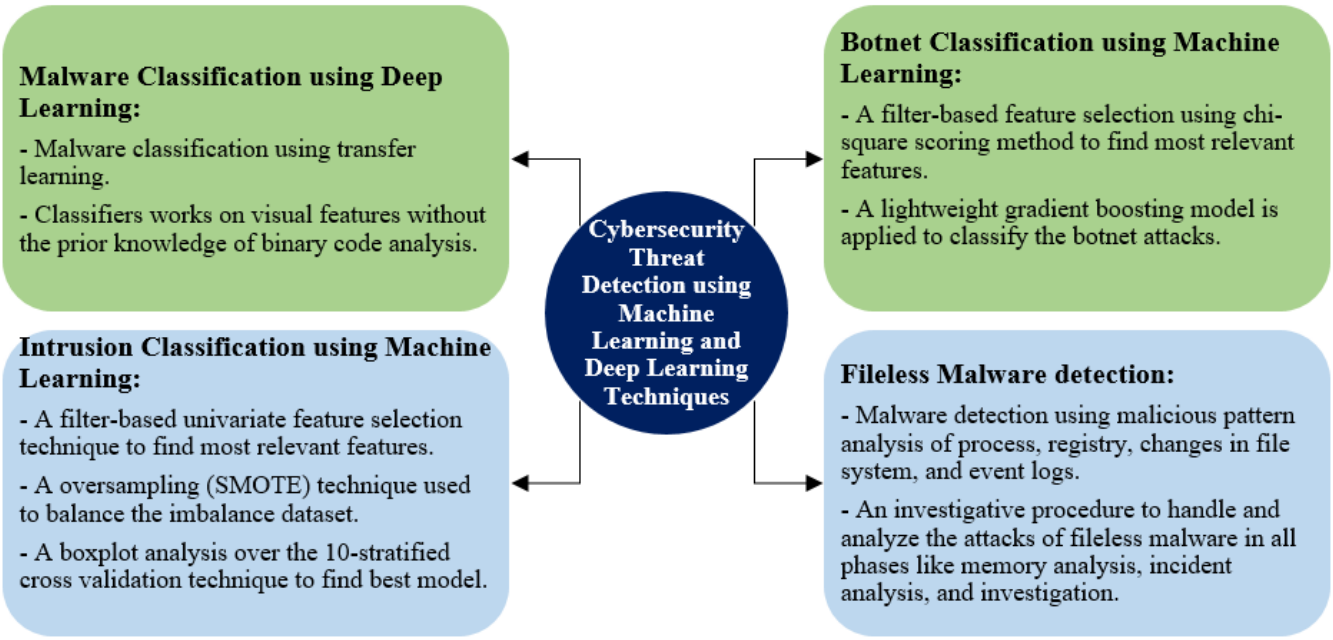

Figure 1. Thesis contributions.

Energy-efficient Computing and Communication. https://doi.org/10.1109/ICRAECC43874.2019.8995028

[12] Sudhakar and Sushil Kumar. 2020. An emerging threat Fileless malware: a survey and research challenges. *Cybersecurity* 3, 1 (2020), 1–12. https://doi.org/10.1186/s42400-019-0043-x

[13] Sudhakar and Sushil Kumar. 2020. ML-IDS: Machine Learning based Intrusion detection System using network-based features. *Microprocessors and Microsystems* (2020). [under review].

[14] Sudhakar and Sushil Kumar. 2021. MCFT-CNN: Malware classification with fine-tune convolution neural networks using traditional and transfer learning in Internet of Things. *Future Generation Computer Systems* 125 (2021), 334–351. https://doi.org/10.1016/j.future.2021.06.029

[15] Danish Vasan, Mamoun Alazab, Sobia Wassan, Hamad Naeem, Babak Safaei, and Qin Zheng. 2020. IMCFN: Image-based malware classification using fine-tuned convolutional neural network architecture. *Computer Networks* 171 (2020), 107138.

[16] Sitalakshmi Venkatraman, Mamoun Alazab, and R Vinayakumar. 2019. A hybrid deep learning image-based analysis for effective malware detection. *Journal of Information Security and Applications* 47 (2019), 377–389.